ChinaXiv合作期刊

樟树叶绿体基因组密码子偏好性分析

秦政[1]，郑永杰[2]，桂丽静[2]，谢谷艾[2]，伍艳芳[2]*

(1. 江西农业大学，林学院，南昌 330045；2 江西省林业科学院，国家林业局樟树树工程技术研究中心 南昌 330032)

摘要：为分析樟树(*Cinnamomum camphora*)叶绿体基因组密码子偏好性使用模式，该研究利用 CodonW、EMBOSS、R 语言等软件和程序，对 53 条樟树叶绿体基因组密码子使用模式及偏好性进行了系统分析。结果表明，樟树叶绿体基因的有效密码子数(ENC)在 36.82~59.30 之间，表明密码子的偏好性较弱。相对同义密码子使用度(RSCU)分析发现 RSCU>1 的密码子有 32 个，其中 28 个以 A、U 结尾，表明第 3 位密码子偏好使用 A 和 U 碱基。中性绘图分析发现，GC_3 与 GC_{12} 的相关性不显著，回归曲线斜率为 0.049，说明密码子偏好性主要受到自然选择的影响，而 ENC-plot 分析中大部分基因落在曲线的下方，同样表明选择是影响密码子偏好性的主要因素。最终共有 9 个密码子(UUU、CUU、UCA、ACA、UAU、AAU、GAU、UGA、GGA)被鉴定为樟树叶绿体基因组的最优密码子。

关键词：樟树，叶绿体基因组，密码子偏好性

中图分类号：Q943.2　　　　文献标识码：A

DOI：10.11931/guihaia.gxzw201805023

Codon Usage Bias Analysis of the Chloroplast Genome of camphor tree (*Cinnamomum camphora*)

QIN Zheng[1], ZHENG Yongjie[2], GUI Lijing[2], XIE Gu’ai[2], WU Yanfang[2]*

(1. *Jiangxi Agricultural University, College of Forestry,* Nanchang 330045, China; 2. *Jiangxi Academy of Forestry, Camphor Engineering and Technology Research Center of State Forestry Administration*, Nanchang 330032, China)

Abstract: Synonymous codons are used with different frequencies among both species and genes within the same genome. In order to analyze the codon usage patterns of the chloroplast genome of camphora tree, 53 CDS(Coding DNA Sequences)were selected from the chloroplast genome of camphora tree and analyzed the codon usage pattern and bias by CodonW, EMBOSS, R language and other softwares. The results showed that the effective codon number (ENC) of the chloroplast gene ranged from 36.82 to 59.30, indicating that bias of codons was weak. There were 32 codons with relative codon usage greater than 1, of which 28 were rich in A and U, indicating that the 3rd position of codon prefers A and U genetic bases. Neutral analysis found that the correlation between GC3 and GC12 was not significant, and the slope of the regression curve was 0.049, suggesting that codon bias was mainly affected by natural selection, while the ENC-plot analysis found that most genes fell below and around the curve. It was shown that the mutation also

---

收稿日期：2018-05-10

基金项目：国家自然科学基金项目（31460209）；江西省林业科学院青年科技人才培养项目（2018522701）[Supported by the National Natural Science Foundation of China（31460209）; Jiangxi Academy of Forestry Youth Science and Technology Talents Training Project（2018522701）]。

作者简介：秦政（1994-），男，江西南昌人，硕士研究生，主要从事植物分子生物学研究，（E-mail）qinzheng_qz@126.com）。

*通信作者：伍艳芳，博士，副研究员，主要从事林木遗传育种研究，（E-mail）yanfangwu2012@163.com。

chinaXiv:201809.00100v1

ChinaXiv合作期刊

affected the formation of codon bias. Eventually，9 codons (UUU, CUU, UCA, ACA, UAU, AAU, GAU, UGA, GGA) were identified as the optimal codons for the chloroplast genome of camphora tree.

Key words *Cinnamoum camphora*, chloroplast genome, codon bias

编码相同氨基酸的不同密码子被称为同义密码子。在生物体中，同义密码子的出现具有非随机性，即某一物种通常倾向于使用一种或几种特定的密码子，该现象被称为密码子使用偏好性(Codon usage bias, CUB)(Grantham et al, 1980; Mar ń et al, 1989)。密码子使用模式的分析从分子生物学层面揭示了基因调控、基因表达、蛋白质二级结构、选择性转录等现象，其中不同生物体密码子使用产生偏好性是突变压力和自然选择下的产生结果(Duret & Mouchiroud, 1999; Gu et al, 2004; Mg & De Farias, 2006)，利用不同物种在密码子使用上的偏好性，可以提高蛋白表达的效率和准确性(Gerrit & Diarmaid, 2016; Lee et al, 2009)。

叶绿体是植物生命活动的代谢中心，在植物光合作用和生物合成中具有重要作用(And and Emes, 2000)。植物叶绿体基因组由于规模小、拷贝数多，且在分子水平上多个基因的功能已得到验证，因此是当今生物学领域的研究热点(Xu et al, 2011)。从 1986 年烟草(*Nicotiana tabacum*)(Shinozaki et al, 1986)的和地钱(Ohyama et al, 1986)(*Marchantia polymorpha*)的叶绿体基因组首次公布以来，目前在 NCBI 网站上的细胞器基因组数据库中中已有包括拟南芥(*Arabidopsis thaliana*)(Sato et al, 1999)、水稻(*Oryza sativa*)(Hiratsuka et al, 1989)、草莓(*Fragaria* × *ananassa*)(Cheng et al, 2017)和连翘(*Forsythia suspensais*)(Wang et al, 2017)等在内的约 800 种植物完整的叶绿体基因组信息。

樟属植物全世界约 250~300 种，我国约有 46 种，主要分布在南方各省区。这些植物不仅可用于提取精油作为化工及医药上的重要原料，同时还是重要家具、建筑及雕刻等用材树种。作为樟属植物中的代表性植物，樟树(*Cinnamomum camphora*)集材用、药用、香料、油用、生态环境建设和生态文化等于一体，极具开发和利用价值。2017 年，Chen 等(Chen et al, 2017)对樟树叶绿体基因组进行了测序和分析，结果显示樟树叶绿体基因组全长 152 570 bp，具有特征性的保守四联体结构，在编码区和非编码区都检测到共 40 个重复结构和 83 个简单序列重复序列。目前，关于樟树叶绿体密码子使用偏好性的研究还未见报道。本研究通过对樟树叶绿体基因组中 53 个基因的密码子使用的偏好性进行分析，在揭示影响密码子偏好性最主要因素的同时还找出了最优密码子，旨在为樟科植物叶绿体基因组学的研究奠定基础。

1 材料与方法

1.1 序列的获取

在 NCBI 数据库(https://www.ncbi.nlm.nih.gov/)中下载樟树及其他樟科植物的叶绿体基因组信息。樟树叶绿体基因组序列的 Genbank 登录号为 MF156716.1，共注释 83 条蛋白质编码基因序列。为避免计算密码子偏好性时出现样本偏差，在剔除重复序列后，最终获得 53 条编码区长度大于 300 bp 且以 ATG 为起始密码子，TAA、TGA 或 TAG 为终止密码子的基因序列用于后续分析。

1.2 方法

1.2.1 密码子偏好性分析

利用 CodonW1.4.2(http://sourceforge.net/projects/codonw)和 Excel 2013 统计各基因的密码子出现次数(Codon number，CN)、有效密码子数(Effective Number of codon, ENC)(Novembre, 2002)和同义密码子相对使用度(Relative Synonymous Condon Usage, RSCU)(Sharp, Li, 1986)。用 EMBOSS(http://imed.med.ucm.es/EMBOSS/)中的 CUSP 程序分析樟树叶绿体基因组中各基因的 GC 总含量和密码子第 1、2、3 位的 GC 含量，分别用 GC_{all}、

chinaXiv:201809.00100v1

chinaXiv:201809.00100v1

GC_1、GC_2 和 GC_3 表示。GC_{3S} 表示同义密码子第 3 位的 GC 含量。樟树叶绿体基因组中各基因的部分参数的相关性及显著性用 R 语言分析并绘图。

1.2.2 中性绘图分析

中性绘图法的作用为初步判断选择或突变对密码子使用偏好性的影响。在 Origin9.1 中，以密码子前两位 GC 含量的平均值 GC_{12} 为纵坐标，GC_3 为横坐标，绘制散点图并做直线拟合分析。在直线拟合分析中，当回归系数接近 1 时， GC_{12} 与 GC_3 间的相关性显著，密码子的碱基组成在 3 个位置上无差异，表明密码子偏好性的主要决定因素为突变；当回归系数接近 0 时，GC_{12} 与 GC_3 之间相关性不显著，说明第密码子 3 个位置上的的碱基组成不同，则密码子的使用更多地受到选择影响。

1.2.3 ENC-plot 绘图分析

ENC-plot 绘图用于分析突变在密码子使用模式中的作用。当实际 ENC 值与预期 ENC 值差异较小时，表明突变对密码子偏好性影响较大，反之选择为影响密码子偏好性的主要因素。此外，还可根据 ENC 值的高低推断密码子偏好性的强弱。ENC 值的理论范围在 21~60，值越小表示偏好性越强。当 ENC 值为 20 和 60 两个极值时，表示同义密码子具有完全偏好性或没有偏好性。用 R 语言绘制二维散点图，横、纵坐标分别为 GC_{3S} 和 ENC，并在图中构建出 ENC 值的期望曲线，曲线的公式如下：

$$ENC = 2+GC_{3S}+29/(GC_{3S}^2+(1-GC_{3S})^2)$$

1.2.4 PR2-plot 分析

为了避免密码子第 3 位的 AT 和 GC 之间突变不平衡，PR2 偏好性分析(PR2-bias plot analysis)仅对 4 种密码子编码的氨基酸做密码子第 3 位上 A、T、C 和 G 这 4 种碱基的组成情况进行分析(Sueoka, 1999)，并以 $A_3/(A_3+T_3)|4$ 的值作为纵坐标，$G_3/(G_3+C_3)|4$ 的值作为横坐标进行绘图。其中，“| 4”表示四密码子氨基酸，分别为丙氨酸，精氨酸(CGA，CGT，CGG，CGC)，甘氨酸，亮氨酸(CTA，CTT，CTG，CTC)，脯氨酸，丝氨酸(TCA，TCT，TCG，TCC)，苏氨酸和缬氨酸。PR2-plot 图的中心点为 A=T，C=G，以该点发出的矢量表示碱基偏移的程度和方向。

1.2.5 最优密码子分析

对前期筛选出的 53 条基因的 ENC 值从小到大进行排序，两端各选出 10%的基因建库，在高低两个基因库中将对应密码子的 RSCU 值相减得到 ΔRSCU，选取 ΔRSCU＞0.08 的密码子作为高表达密码子。将前文分析中 RSCU 值大于 1 的高频率密码子与 ΔRSCU＞0.08 高表达密码子的结果相结合，从而确定出最优密码子(胡莎莎等, 2016; 罗洪等, 2015)。

1.2.6 系统进化树的构建与分析

使用 MAFFT 7.397(Katoh, Standley, 2013)(https://mafft.cbrc.jp/alignment/software/)将樟树与其他八种樟科植物的叶绿体基因组序列做多序列比对分析。选择银杏和云杉作为外类群。用 BioEdit 手工校正比对结果后导入 MEGA7.0 中用系统邻接法(Neighbor-Joining, NJ)构建进化树。用 Bootstrap method 进行重复检验 1 000 次得出结果。

2 结果与分析

2.1 樟树叶绿体基因组密码子偏好性分析

利用 CodonW1.4.2 分析樟树叶绿体基因组中 53 条候选基因的密码子组成 (表 1)。结果显示，平均 GC 含量为 39.1%，且分布趋势为 GC_1(46.81%)>GC_2(38.77%)>GC_3(31.77%)，表明在密码子的 3 个位置上 GC 并不是均匀地分布，且第 3 位密码子偏好使用 A 和 U 碱基。樟树叶绿体基因组 ENC 的取值范围在 36.82~59.3，且大部分在 45 以上，说明樟树叶绿体基因组密码子偏好性较弱。

图 1 结果表明，RSCU 值大于 1 的密码子数目为 34 个。其中第 3 位上的碱基为 U、A

和 G 的数目分别为 16，13 和 2，这说明 U 和 A 为樟树叶绿体基因组密码子的偏好碱基。53 条基因的 GC 含量、ENC 和 CN 间相关性分析结果显示(表 2)，GC_{all} 与 GC_1、GC_2 和 GC_3 呈极显著相关，但 GC_1、GC_2 和 GC_3 的相关性水平不显著。ENC 与 GC_3 相关系数为 0.27，呈显著相关，说明密码子的使用偏好性受第 3 位碱基的影响。CN 与 GC_3、ENC 显著相关，表明 GC_3 和密码子的偏好性受基因序列长度的影响。

2.2 中性绘图分析

中性绘图分析显示(图 2)，GC_{12} 和 GC_3 的分布范围都较为集中，GC_{12} 的范围在 0.3431~0.5458，GC_3 在 0.2244~0.5214。GC_{12} 与 GC_3 的相关系数为 0.143，回归系数为 0.049，说明两者的相关性不显著，密码子第 3 位与第 1、2 位的进化方式可能存在一定差异，因此樟树叶绿体基因密码子偏好性主要受到选择的影响。

2.3 ENC-plot 绘图

绘制 ENC-plot 图分析樟树叶绿体基因组密码子偏好性是否被突变或选择等因素影响。从图 3 可明显看出，大部分基因落在期望曲线的下方，表示 ENC 实际值与 ENC 预期值之间差异较大，也说明樟树叶绿体基因组密码子偏好性更多受选择的影响。

表 1 樟树叶绿体基因组各基因密码子不同位置的 GC 含量

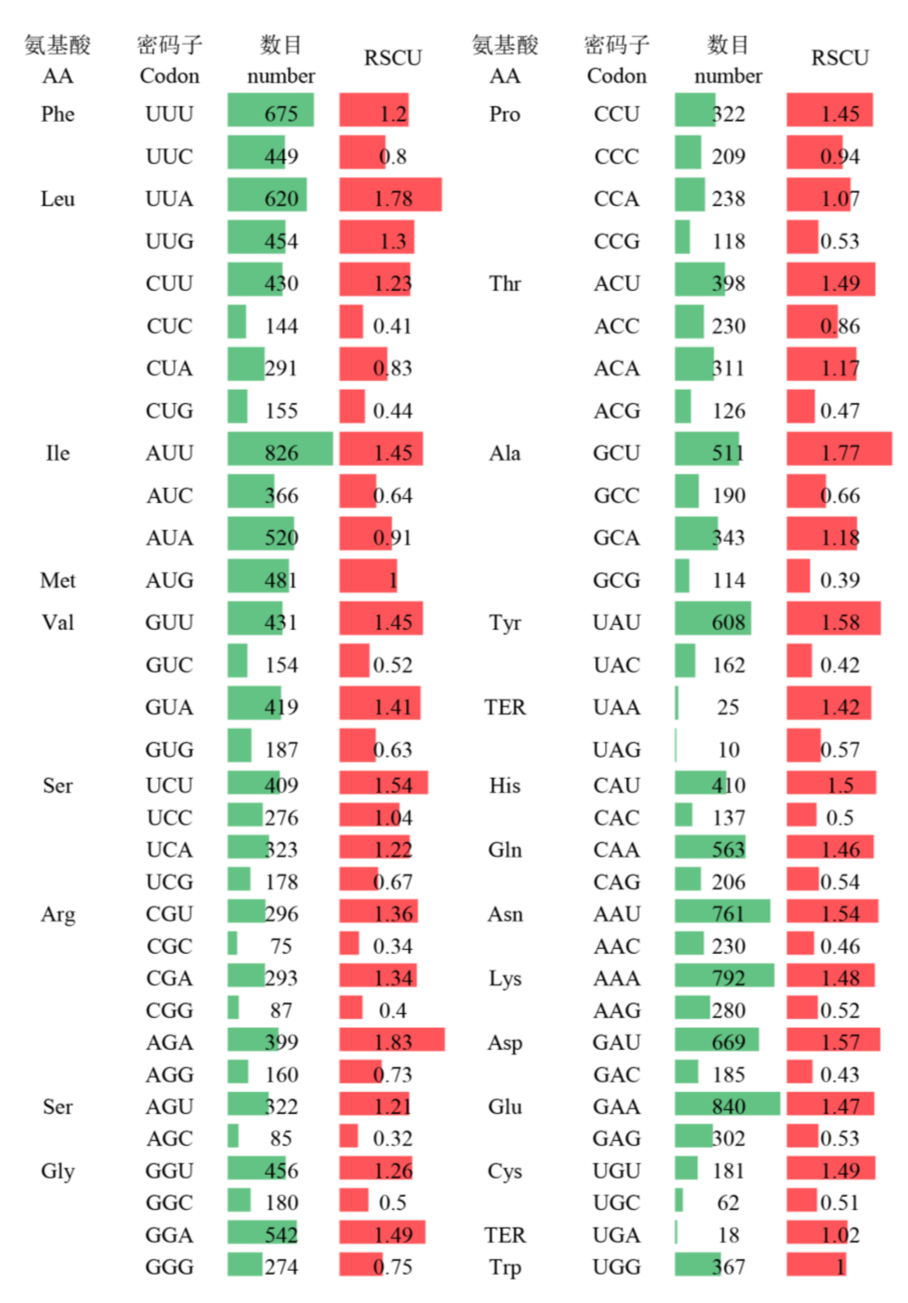
Table 1 GC content of different position of each gene in the chloroplast genome of *C. camphora*

基因 Gene	GC 含量				ENC	基因 Gene	GC 含量				ENC
	GC_{all}	GC_1	GC_2	GC_3			GC_{all}	GC_1	GC_2	GC_3	
rps12	41.94	51.61	48.39	25.81	42.31	*rpl20*	40.11	39.83	51.69	28.81	47.04
psbA	42.28	49.72	43.79	33.33	42.62	*rps12*	41.46	47.90	48.74	27.73	45.19
matK	36.18	43.88	32.62	32.04	52.93	*clpP*	43.07	58.91	37.62	32.67	55.32
atpA	42.39	56.30	40.55	30.31	49.82	*psbB*	44.86	54.81	45.97	33.79	52.93
atpF	39.28	51.35	35.14	31.35	50.25	*petB*	40.90	49.54	41.67	31.48	45.28
atpI	39.25	50.00	37.50	30.24	46.54	*petD*	40.08	51.53	39.26	29.45	44.72
rps2	39.38	41.35	43.46	33.33	52.53	*rpoA*	36.67	46.47	34.12	29.41	49.35
rpoC2	38.72	46.61	38.31	31.24	51.18	*rps11*	44.53	54.20	54.96	24.43	49.29
rpoC1	39.67	51.02	39.04	28.95	50.58	*rps8*	36.59	42.11	40.60	27.07	48.44
rpoB	40.56	50.33	39.30	32.05	51.43	*rpl14*	40.38	55.28	36.59	29.27	44.71
psbD	42.84	51.69	43.50	33.33	45.37	*rpl16*	44.44	51.11	54.81	27.41	36.98
psbC	44.80	54.22	45.99	34.18	47.31	*rps3*	37.25	50.23	33.48	28.05	53.47
rps14	42.24	43.56	47.52	35.64	37.89	*rpl22*	38.12	41.49	42.55	30.32	50.08
psaB	42.27	49.12	42.72	34.97	51.27	*ycf2*	37.38	41.27	33.82	37.05	53.72
psaA	43.63	53.00	43.54	34.35	51.09	*rps7*	39.96	53.21	44.23	22.44	47.57
ycf3	39.05	47.34	40.24	29.59	54.44	*ycf68*	50.43	47.86	51.28	52.14	47.88
rps4	39.11	50.00	40.59	26.73	51.35	*ycf1*	32.75	38.31	30.31	29.62	49.50
ndhJ	42.14	51.57	38.36	36.48	54.1	*ndhH*	39.68	52.03	36.29	30.71	51.94
ndhK	38.81	42.31	43.71	30.42	50.81	*ndhA*	37.86	46.20	40.76	26.63	45.48
ndhC	38.29	49.59	34.71	30.58	47.37	*ndhI*	35.73	39.78	39.23	28.18	51.4

chinaXiv:201809.00100v1

ChinaXiv合作期刊

atpE	41.73	50.37	40.74	34.07	50.24	*ndhG*	37.29	46.89	35.59	29.38	46.18
atpB	43.02	56.91	41.68	30.46	48.63	*ndhE*	34.31	40.20	35.29	27.45	48.23
rbcL	45.73	58.61	43.91	34.66	51.37	*ccsA*	34.60	34.70	40.06	29.02	50.14
accD	37.04	41.27	39.88	29.96	47.01	*ndhF*	34.81	38.26	37.99	28.19	47.90
ycf4	40.72	43.78	42.16	36.22	59.30	*ycf1*	34.90	38.78	31.63	34.29	53.62
petA	41.12	53.27	36.14	33.96	54.99	*rps7*	39.96	53.21	44.23	22.44	47.57
rps18	34.64	35.29	43.14	25.49	36.82	*Total*	40.62	49.04	40.99	31.84	49.24



氨基酸 AA	密码子 Codon	数目 number	RSCU	氨基酸 AA	密码子 Codon	数目 number	RSCU
Phe	UUU	675	1.2	Pro	CCU	322	1.45
	UUC	449	0.8		CCC	209	0.94
Leu	UUA	620	1.78		CCA	238	1.07
	UUG	454	1.3		CCG	118	0.53
	CUU	430	1.23	Thr	ACU	398	1.49
	CUC	144	0.41		ACC	230	0.86
	CUA	291	0.83		ACA	311	1.17
	CUG	155	0.44		ACG	126	0.47
Ile	AUU	826	1.45	Ala	GCU	511	1.77
	AUC	366	0.64		GCC	190	0.66
	AUA	520	0.91		GCA	343	1.18
Met	AUG	481	1		GCG	114	0.39
Val	GUU	431	1.45	Tyr	UAU	608	1.58
	GUC	154	0.52		UAC	162	0.42
	GUA	419	1.41	TER	UAA	25	1.42
	GUG	187	0.63		UAG	10	0.57
Ser	UCU	409	1.54	His	CAU	410	1.5
	UCC	276	1.04		CAC	137	0.5
	UCA	323	1.22	Gln	CAA	563	1.46
	UCG	178	0.67		CAG	206	0.54
Arg	CGU	296	1.36	Asn	AAU	761	1.54
	CGC	75	0.34		AAC	230	0.46
	CGA	293	1.34	Lys	AAA	792	1.48
	CGG	87	0.4		AAG	280	0.52
	AGA	399	1.83	Asp	GAU	669	1.57
	AGG	160	0.73		GAC	185	0.43
Ser	AGU	322	1.21	Glu	GAA	840	1.47
	AGC	85	0.32		GAG	302	0.53
Gly	GGU	456	1.26	Cys	UGU	181	1.49
	GGC	180	0.5		UGC	62	0.51
	GGA	542	1.49	TER	UGA	18	1.02
	GGG	274	0.75	Trp	UGG	367	1

注：Phe. 苯丙氨酸；Leu. 亮氨酸；Ile. 异亮氨酸；Met. 蛋氨酸；Val，缬氨酸；Ser. 丝氨酸；Arg. 精氨酸；Gly.甘氨酸；Pro. 脯氨酸；Thr. 苏氨酸；Ala. 丙氨酸；Tyr. 酪氨酸；His. 组氨酸；Gln. 谷氨酰胺；Asn. 天冬酰胺；Lys. 赖氨酸；Asp. 天冬氨酸 ；Glu. 谷氨酸；Cys. 半胱氨酸；Trp. 色氨酸；TER 表示有空白或内部终止符号的氨基酸。下同。

chinaXiv:201809.00100v1

ChinaXiv合作期刊

Note: Phe. Phenylalanine; Leu. Leucine; Ile. Isoleucine; Met. Methionine; Val. Valine; Ser. Serine; Arg. Argnine; Gly. Glycine; Pro. Proline; Thr. Threonine; Ala. Alanine; Tyr. Tyrosine; Hls. Hlstidine; Gln. Glutamine; Asn. Asparagine; Lys. Lysine; Asp. Aspartic acid; Glu. Glutamic acid; Cys. Cysteine; Trp. Tryptophane; TER indicates an amino acid with a blank or internal termination symbol. The same below.

图 1 樟树叶绿体基因组同义密码子使用频率

Fig.1 Relative synonymous codons usage analysis in the chloroplast of *C.camphora*

表 2 樟树叶绿体基因组中各参数的相关性

Table 2 Correlation analysis of each parameters in the chloroplast genome of *C.camphora*

项目 Item	G_{call}	GC1	GC2	GC3	ENC	CN
Gcall	1.000					
GC1	0.722**	1.000				
GC2	0.662**	0.193	1.000			
GC3	0.487**	0.058	0.020	1.000		
ENC	-0.066	0.074	-0.428**	0.274*	1.000	
CN	-0.207	-0.165	0.355	0.186	0.262*	1.000

注：*表示相关性达到显著水平(P<0.05)，**表示相关性达到极显著水平(P<0.01)。

Note：* stands for significant correlation at the 0.05 level；

** stands for significant correlation at the 0.01 level.

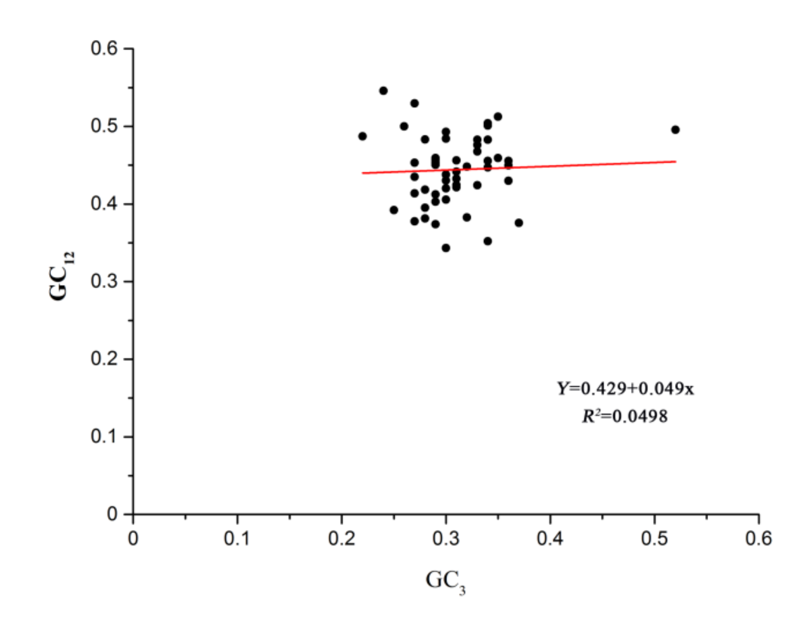


图 2 樟树叶绿体基因组中性绘图分析

Fig.2 Neutrality plot analysis in the chloroplast of *C.camphora*

chinaXiv:201809.00100v1

ChinaXiv合作期刊

chinaXiv:201809.00100v1

表 2 樟树叶绿体基因组最优密码子分析

Table 2 GC content of different position of genes in the chloroplast genome of *C.camphora*

氨基酸 Amino acid	密码子 Codon	高表达基因 High expressed gene Number	RSCU	低表达基因 Low expressed gene Number	RSCU	Δ RSCU
Phe	UUU *	28	1.08	16	0.84	0.24
	UUC	24	0.92	22	1.16	-0.24
Leu	UUA	19	1.34	19	1.84	-0.5
	UUG	19	1.34	17	1.65	-0.31
	CUU *	23	1.62	15	1.45	0.17
	CUC *	6	0.42	0	0	0.42
	CUA	9	0.64	10	0.97	-0.33
	CUG ***	9	0.64	1	0.1	0.54
Ile	AUU	34	1.15	33	1.52	-0.37
	AUC *	26	0.88	16	0.74	0.14
	AUA *	29	0.98	16	0.74	0.24
Met	AUG	26	1	22	1	0
Val	GUU	28	1.53	18	1.53	0
	GUC	6	0.33	6	0.51	-0.18
	GUA	23	1.26	20	1.7	-0.44
	GUG ***	16	0.88	3	0.26	0.62
Ser	UCU	20	1.82	21	2	-0.18
	UCC	10	0.91	12	1.14	-0.23
	UCA **	11	1	7	0.67	0.33
	UCG	5	0.45	5	0.48	-0.03
Pro	CCU	17	1.15	18	1.67	-0.52
	CCC *	21	1.42	13	1.21	0.21
	CCA	16	1.08	11	1.02	0.06
	CCG *	5	0.34	1	0.09	0.25
Thr	ACU	11	1.16	19	1.65	-0.49
	ACC	11	1.16	15	1.3	-0.14
	ACA *	10	1.05	11	0.96	0.09
	ACG ***	6	0.63	1	0.09	0.54
Ala	GCU	27	1.66	32	2.21	-0.55
	GCC **	12	0.74	6	0.41	0.33
	GCA	21	1.29	18	1.24	0.05
	GCG *	5	0.31	2	0.14	0.17
Tyr	UAU **	45	1.76	17	1.42	0.34
	UAC	6	0.24	7	0.58	-0.34
TER	UAA	2	1.2	4	2.4	-1.2
	UAG ***	1	0.6	0	0	0.6
His	CAU	11	1.38	15	1.36	0.02
	CAC	5	0.63	7	0.64	-0.01

Gln	CAA	30	1.36	19	1.36	0
	CAG	14	0.64	9	0.64	0
Asn	AAU ***	38	1.73	20	1.18	0.55
	AAC	6	0.27	14	0.82	-0.55
Lys	AAA	25	1.28	29	1.53	-0.25
	AAG *	14	0.72	9	0.47	0.25
Asp	GAU **	37	1.72	11	1.38	0.34
	GAC	6	0.28	5	0.63	-0.35
Glu	GAA	54	1.5	28	1.6	-0.1
	GAG	18	0.5	7	0.4	0.1
Cys	UGU	5	1	7	1.75	-0.75
	UGC ***	5	1	1	0.25	0.75
TER	UGA ***	2	1.2	1	0.6	0.6
Trp	UGG	18	1	15	1	0
Arg	CGU	14	1.25	22	1.67	-0.42
	CGC *	5	0.45	4	0.3	0.15
	CGA	15	1.34	18	1.37	-0.03
	CGG **	6	0.54	2	0.15	0.39
	AGA	16	1.43	27	2.05	-0.62
	AGG ***	11	0.99	6	0.46	0.53
Ser	AGU	14	1.27	13	1.24	0.03
	AGC	6	0.55	5	0.48	0.07
Gly	GGU	22	1.19	29	1.71	-0.52
	GGC	11	0.59	9	0.53	0.06
	GGA *	30	1.62	23	1.35	0.27
	GGG *	11	0.59	7	0.41	0.18

注：*表示最优密码子。

Note：*represents the optimal codons.

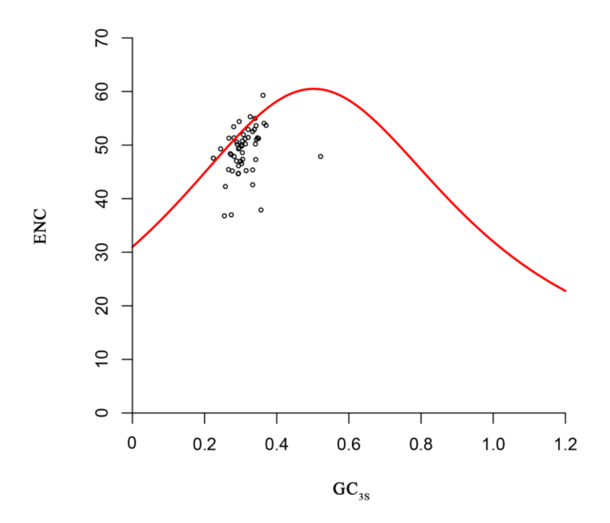


图 3 樟树叶绿体基因组 ENC-plot 绘图分析

Fig.3 Analysis of ENC-plot in the chloroplast of *C.camphora*

chinaXiv:201809.00100v1

ChinaXiv合作期刊

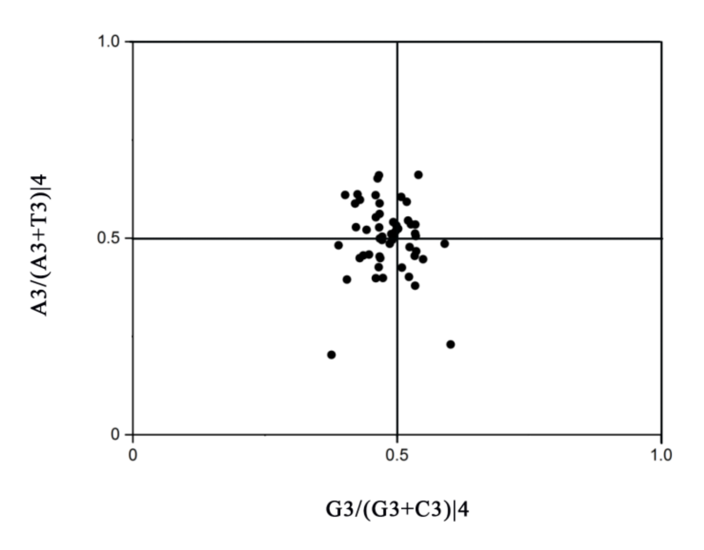
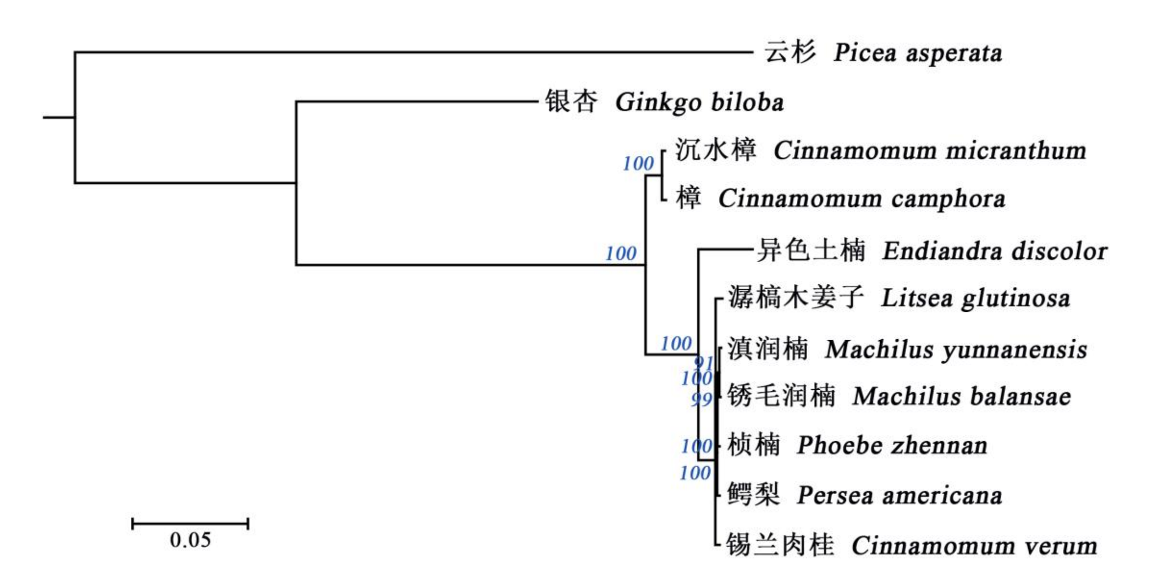


图 4 樟树叶绿体基因组 PR2-plot 分析

Fig.4 Analysis of PR2-plot in the chloroplast of *C.camphora*

chinaXiv:201809.00100v1



注：NCBI 登录号：潺槁木姜子 KU382356.1；沉水樟 NC_035802.1；滇润楠 KT348516.1；鳄梨 KX437771.1；锡兰肉桂 NC_035236.1；锈毛润楠 KT348517.1；异色土楠 KX674308.1；银杏 JN867585.1；云杉 NC_032367.1；樟树 MF156716.1；桢楠 NC_036143.1。

Note：ID number in NCBI: *Litsea glutinosa* KU382356.1; *Cinnamomum micranthum* NC_035802.1; *Machilus yunnanensis* KT348516.1; *Persea americana* KX437771.1; *Cinnamomum verum* NC_035236.1; *Machilus balansae* KT348517.1; *Endiandra discolor* KX674308.1; *Ginkgo biloba* JN867585.1; *Picea asperata* NC_032367.1; *Cinnamomum camphora* MF156716.1; *Phoebe zhennan* NC_036143.1).

图 5 樟科叶绿体基因组系统进化树

Fig.5 Phylogenetic tree based on chloroplast genome sequence of *Lauraceae*

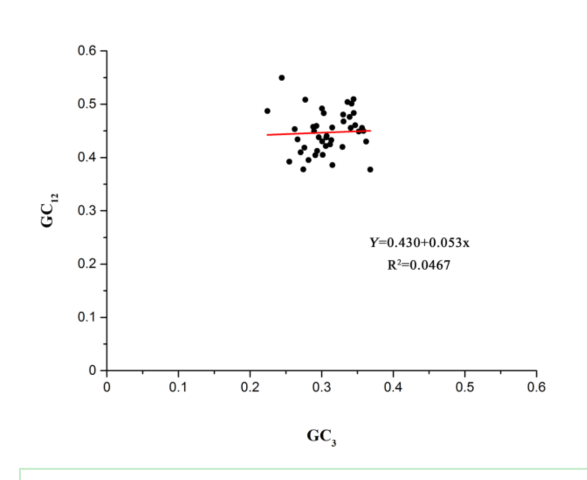


图 6 沉水樟叶绿体基因组中性绘图分析
Fig.7 Analysis of neutrality plot of *C.micranthum*

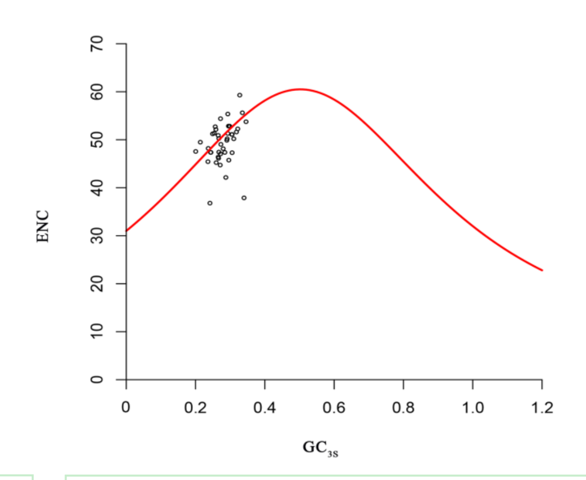


图 7 沉水樟叶绿体基因组 ENC-plot 分析
Fig.8 Analysis of ENC-plot of *C.micranthum*

chinaXiv:201809.00100v1

2.4 PR2-plot 分析

通过 PR2-plot 绘图进一步分析樟树叶绿体基因组密码子偏好性的影响(图 4)。当 A、T、C 和 G 四种碱基的使用频率，即 A=T，C=G 时，密码子使用偏好性不受选择和突变影响(Sueoka, 2001)。研究发现，4 个区域中的基因位点分布并不均匀，左上方位置的个数较多，说明在第 3 位碱基的选择上密码子具有偏好性，且碱基 A 的使用频率高于 T，碱基 C 的使用频率高于 G。因此可推论选择是樟树叶绿体基因组密码子偏好性的主要影响因素。

2.5 最优密码子的确定

对 53 个候选基因的 ENC 值进行排序后，从两端各选取 10%的基因建立高、低表达基因库，然后计算两个表达库中密码子的 RSCU 值和两个库中的 ΔRSCU 值(表 2)。最后确定了 26 个密码子为樟树叶绿体基因组高表达密码子。对 26 个密码子的第 3 位碱基进行分析，其中 U 结尾的有 5 个，以 A 结尾的有 5 个，以 C、G 结尾的分别为 6 个和 10 个。结合 26 个高表达密码子和图 1 的 32 个高频密码子，选取两者的公共部分，最后确定了 9 个樟树叶绿体基因组最优密码子，分别为 UUU、CUU、UCA、ACA、UAU、AAU、GAU、UGA、GGA。其中，所有密码子均以 U 或 A 结尾。

2.6 系统进化树的构建与分析

比较叶绿体基因组序列可以揭示近缘物种类群之间的系统进化关系，同时阐明植物的进化模式(Kim et al., 2015)。为深入了解樟树在樟科植物的进化规律和关系，将樟树与其他八种樟科植物的叶绿体基因组序列共同构建系统进化树，选取裸子植物的银杏和云杉为外类群(图 5)。结果显示，樟科植物的叶绿体基因组单独聚为一枝，与外类群区分明显。樟树与沉水樟的叶绿体基因组亲缘关系最近，且两者均为樟科樟属植物成员，因此推测可能两者可能具有相似的密码子偏好性模式。利用相同的方法对沉水樟叶绿体基因组进行中性绘图(图 6)和 ENC-plot 绘图(图 7)分析，发现 GC_{12} 与 GC_3 的回归系数接近 0(为 0.053)，同样在 ENC-plot 绘图中，大部分基因落在期望曲线的下方，两者都说明选择是影响沉水樟叶绿体基因组密码子偏好性的主要因素，这也与樟树的分析结果类似，表明樟科或樟属植物在进化过程中自然选择因素对叶绿体基因组密码子偏好性的影响较大。

chinaXiv:201809.00100v1

3 讨论

同义密码子在不同物种间和同一基因组内的不同基因间使用都具有一定偏好性，高表达量的基因通常具有最优密码子并且其密码子偏好性往往更强(Ghaemmaghami et al, 2003; Goetz & Fuglsang, 2005; Ingvarsson, 2007)。Zhou(Zhou et al, 2008b)等的研究发现，植物叶绿体基因组中，密码子结尾处的碱基更偏好使用 A 或 T。本研究发现在樟树叶绿体基因组中 GC_3 与 GC_1、GC_2 相关不显著，且 GC_3 的比例为三者中最低，因此也得到类似的结果，即偏好密码子多以 A 或 T 结尾，这也与苹果(*Malus* × *domestica*)(金桂花等, 2014)、杉木(*Cunninghamia lanceolata*)(郑薇玮等, 2016)和柿(*Diospyros kaki*)(傅建敏等, 2017)等植物的密码子偏好性研究结果一致。

密码子偏好性的形成原因受多种因素影响，除了最主要的突变和自然选择(Rao et al, 2011)外，碱基组成差异(Romero et al, 2000)、基因编码结构(Rao et al, 2011)、tRNA 丰度(Novoa & Pouplana, 2012)等多种因素也会影响其偏好性。其中 tRNA 的表达丰度最高，相应的密码子偏好性就越强(Duret, 2002; Hershberg & Petrov, 2008)。本研究通过对樟树叶绿体基因组三个不同位置上的密码子碱基组成及密码子相关系数进行研究，经中性分析、ENC-plot 分析和 PR2-plot 等分析后发现，选择是影响樟树叶绿体基因组密码子偏好性的主要因素，突变对密码子的偏好性影响较小。在陆地棉(*Gossypium hirsutum*)(尚明照等, 2011)、糜子(*Panicum miliaceum*)(刘慧等, 2017)、普通油茶(*Camellia oleifera*)(王鹏良等, 2018)等植物的叶绿体基因组密码子偏好性分析中，也得到了与本研究相似的结果。然而，在对拟南芥和杨树(Zhou et al, 2008)叶绿体基因组密码子偏好性的研究中发现，突变是主要的影响因素。由此可见植物密码子偏好性受多重因素共同影响，且影响不同植物密码子偏好性的主要因素也不同。

在强的正向选择和突变压力的作用下往往会形成大量的最优密码子，相反在纯化选择和突变压力作用下最优密码子数目相对较少(Hershberg & Petrov, 2008; 宋辉等, 2015)。本研究将高频密码子和高表达密码子分析相结合，共筛选出 9 个樟树叶绿体基因组中的最优密码子，且这些密码子都以 A 和 U 碱基结尾，该结果与前人在蒺藜苜蓿(杨国锋等, 2015)、水稻(Zhou et al, 2008a)等植物的中的研究相吻合。在系统进化树中，分枝越长的物种表明其进化速率与树中的其他成员相比更快，进化树中异色土楠自成一枝且分枝相对较长，这也与前人发现它是在进化中首先与其他樟科植物分离的物种的观点相吻合(Hinsinger and Strijk, 2017)。同时，通过分析发现樟树与沉水樟的密码子偏好性都主要受到选择的影响，说明亲缘关系越近的物种可能具有越相似的密码子偏好性。本研究对今后提高目的基因的表达效率具有重要作用，而樟科植物叶绿体基因组密码子分析也将为今后被子植物的系统发育研究提供重要的参考价值(Moore et al, 2007; Ruhfel et al, 2014)。

ChinaXiv合作期刊

chinaXiv:201809.00100v1

参考文献

CHEN C, ZHENG Y, LIU S, et al, 2017. The complete chloroplast genome of *Cinnamomum camphora* and its comparison with related Lauraceae species[J]. Peer J, 5(2): e3820.

CHENG H, LI J, ZHANG H, et al, 2017. The complete chloroplast genome sequence of strawberry (*Fragaria* × *ananassa Duch.*) and comparison with related species of Rosaceae[J]. PeerJ, 5(10): e3919.

DURET L, 2002. Evolution of synonymous codon usage in metazoans[J]. Curr Opin Genet Dev, 12(6): 640-649.

DURET L＆MOUCHIROUD D, 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis[J]. Proc Natl Acad Sci USA, 96(8): 4482-4487.

FU JM, SUO YJ, LIU HM, et al, 2017. Analysis on codon usage in the chloroplast protein-coding genes of *Diospyros* spp[J]. Economic For Res, 35(2): 38-44.[傅建敏, 索玉静, 刘慧敏, 等, 2017. 柿属植物叶绿体蛋白质编码基因密码子用法[J]. 经济林研究, 35(2): 38-44.]

GERRIT B＆DIARMAID H, 2016. The selective advantage of synonymous codon usage bias in Salmonella[J]. Plos Genet, 12(3): e1005926.

GHAEMMAGHAMI S, HUH WK, BOWER K, et al, 2003. Global analysis of protein expression in yeast[J]. Nature, 425(6959): 737-741.

GOETZ R＆FUGLSANG A, 2005. Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from Escherichia coli[J]. Biochem Bioph Res Co, 327(1): 4-7.

GRANTHAM R, GAUTIER C, GOUY M, et al, 1980. Codon catalog usage and the genome hypothesis[J]. Nucleic Acids Res, 8(1): r49.

GU W, ZHOU T, MA J, et al, 2004. Analysis of synonymous codon usage in SARS Coronavirus and other viruses in the Nidovirales[J]. Virus Res, 101(2): 155-161.

HERSHBERG R＆PETROV DA, 2008. Selection on codon bias[J]. Annu Rev Genet, 42(42): 287-299.

HINSINGER DD＆STRIJK JS, 2017. Toward phylogenomics of Lauraceae: The complete chloroplast genome sequence of *Litsea glutinosa* (Lauraceae), an invasive tree species on Indian and Pacific Ocean islands[J]. Plant Gene, 9: 71-79.

HIRATSUKA J, SHIMADA H, WHITTIER R, et al, 1989. The complete sequence of the rice (*Oryza sativa*) chloroplast genome: Intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals[J]. Mol Gen Genet, 217(2-3): 185-194.

HU SS, LUO H, WU Q, et al, 2016. Analysis of codon bias of chloroplast genome of tartary buckwheat[J]. Molecular Plant Breeding, 14(2): 309-317.[胡莎莎, 罗洪, 吴琦, 等, 2016. 苦荞叶绿体基因组密码子偏爱性分析[J]. 分子植物育种, 14(2): 309-317.]

INGVARSSON PK, 2007. Gene expression and protein length influence codon usage and rates of sequence evolution in *Populus tremula*[J]. Mol biol evol, 24(3): 836.

JIN GH, CHEN SY, YIN TS, et al, 2014. Characterization of the complete chloroplast genome of apple (*Malus* × *domestica*，Rosaceae) [J]. J Plant Classification Resource, 36(4): 468-484.[金桂花, 陈斯云, 伊廷双, 等. 2014. 苹果叶绿体基因组特征分析[J]. 植物分类与资源学报, 36(4): 468-484.]

KATOH K＆STANDLEY DM, 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability[J]. Mol Biol Evol, 30(4): 772-780.

ChinaXiv合作期刊

KIM K, LEE SC, LEE J, et al, 2015. Complete chloroplast and ribosomal sequences for 30 accessions elucidate evolution of Oryza AA genome species[J]. Sci Rep-uk, 5: 15655.
LEE SF, LI YJ, HALPERIN S A, 2009. Overcoming codon-usage bias in heterologous protein expression in Streptococcus gordonii[J]. Microbiology, 155(Pt 11): 3581-3588.
LIU H, WANG MX, YUE WJ, et al, 2017. Analysis of codon usage in the chloroplast genome of Broomcornillet (*Panicum miliaceum* L.) [J]. J Plant Sci, 35(3): 2-371.[刘慧，王梦醒，岳文杰，等, 2017. 糜子叶绿体基因组密码子使用偏性的分析[J]. 植物科学学报, 35(3): 362-371.]
LUO H, WU SS, WU Q, et al, 2015. Analysis of buckwheat chloroplast gene codon bias[J]. Gen Appl Biol, 34(11): 2457-2464.[罗洪，胡莎莎，吴琦，等，2015. 甜荞叶绿体基因密码子偏爱性分析[J]. 基因组学与应用生物学, 34(11): 2457-2464.]
MARÍN A, BERTRANPETIT J, OLIVER JL, et al, 1989. Variation in G+C content and codon choice: differences among synonymous codon groups in vertebrate genes[J]. Nucleic Acids Res, 17(15): 6181-6189.
MG VDL＆FARIAS ST, 2006. Correlation between codon usage and thermostability[J]. Extremophiles, 10(5): 479-481.
MOORE MJ, BELL CD, SOLTIS PS, et al, 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms[J]. Proceed Nat Acad Sci USA, 104(49): 19363-19368.
NEUHAUS HE＆EMES MJ, 2000. Nonphotosynthetic metabolism in plastids[J]. Annu. Rev. Plant Physiol. Plant Mol Biol, 51(1):111-140
NOVEMBRE JA, 2002. Accounting for background nucleotide composition when measuring codon usage bias[J]. Mol Biol Evol, 19(8): 1390.
NOVOA EM，POUPLANA LRD, 2012. Speeding with control: codon usage, tRNAs, and ribosomes[J]. Trends Genet, 28(11): 574.
OHYAMA K, FUKUZAWA H, KOHCH T, et al, 1986. Chloroplast gene organization deduced from complete sequence of liverwort Marchantia polymorpha chloroplast DNA[J]. Nature, 322(6079): 572-574.
RAO Y, WU G, WANG Z, et al. 2011. Mutation bias is the driving force of codon usage in the gallus gallus genome[J]. DNA Res, 18(6): 499-512.
ROMERO H, ZAVALA A, MUSTO H, 2000. Codon usage in chlamydia trachomatis is the result of strand-specific mutational biases and a complex pattern of selective forces[J]. Nucl Acids Res, 28(10): 2084-2090.
RUHFEL BR, GITZENDANNER MA, SOLTIS PS, et al, 2014. From algae to angiosperms–inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes[J]. Bmc evol biol, 14(1): 23.
SATO S, NAKAMURA Y, KANEKO T, et al, 1999. Complete structure of the chloroplast genome of *Arabidopsis thaliana*[J]. DNA Res, 6(5): 283.
SHANG ZM, LIU F, HUA JP, et al, 2011. Analysis on codon usage of chloroplast genome of *Gossypium hirsutum* [J]. Chin J Agric Sci, 44(2): 245-253.[尚明照，刘方，华金平，等, 2011. 陆地棉叶绿体基因组密码子使用偏性的分析[J]. 中国农业科学, 44(2): 245-253.]
SHARP PM＆LI WH, 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms[J]. J Mol Evol, 24(1-2): 28-38.
SHINOZAKI K, OHME M, TANAKA M, et al, 1986. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression[J]. Embo J, 5(9): 111-148.

chinaXiv:201809.00100v1

SONG H, WANG PF, MA DC, et al, 2015. Analysis on codon usage of WRKY transcription factor of *Medicago truncatula*[J]. J Agric Biotechnol, 23(2): 203-212.[宋辉，王鹏飞，马登超，等，2015. 蒺藜苜蓿 WRKY 转录因子密码子使用偏好性分析[J]. 农业生物技术学报，23(2): 203-212.]

SUEOKA N. 1999. Translation-coupled violation of Parity Rule 2 in human genes is not the cause of heterogeneity of the DNA G+C content of third codon position[J]. Gene, 238(1): 53-58.

SUEOKA N. 2001. Near homogeneity of pr2-bias fingerprints in the human genome and their implications in phylogenetic analyses[J]. J Mol Evol, 53(4-5): 469.

WANG PL, YANG LP, WU HY, et al, 2018. Condon preference of chloroplast genome in *Camellia oleifera*[J]. Guangxi Plant (2).[王鹏良，杨利平，吴红英，等, 2018. 普通油茶叶绿体基因组密码子偏好性分析[J]. 广西植物(2).]

WANG W, YU H, WANG J, et al. 2017. The complete chloroplast genome sequences of the medicinal plant *Forsythia suspensa* (Oleaceae) [J]. Int J Mol Sci, 18(11): 2288.

XU C, CAI X, CHEN Q, et al. 2011. Factors affecting synonymous codon usage bias in chloroplast genome of *Oncidium Gower Ramsey*[J]. Evol bioinform, (7): 271-278.

XU C, BEN AL, CAI XN, et al. 2010. Analysis of synonymous codon usage in chloroplast genome of phalaenopsis aphrodite subsp.formosana[J]. Mol Plant Breeding, 2010, 08(5): 945-950.[续晨，贲爱玲，蔡晓宁,等，2010.蝴蝶兰叶绿体基因组密码子使用的相关分析[J]. 分子植物育种, 08(5):945-950.]

YANG GF, SU KL, ZHAO YR, et al. Analysis on codon usage of chloroplast genome of *Medicago truncatula* [J]. J Pratac Sin, 24(12): 171-179.[杨国锋，苏昆龙，赵怡然，等, 2015. 蒺藜苜蓿叶绿体密码子偏好性分析[J]. 草业学报, 24(12): 171-179.]

ZHANG X, XIN W, WANG S, et al, 2005. Optimizing the codon usage of HIV-1 gag gene according to the codon bias of vaccinia virus improves the gag gene expression[J]. Chinese Journal of Virology, 21(3): 210-216.

ZHENG WW, CHEN JH, HAO ZD, et al, 2016. Analysis on codon usage bias of chloroplast genes from five conifers including *Cunninghamia lanceolata* [J]. Mol Plant Breeding, 14(5): 1091-1097.[郑薇玮，陈金慧，郝兆东，等，2016. 杉木等 5 种针叶树叶绿体密码子偏好性分析[J]. 分子植物育种, 14(5): 1091-1097.]

ZHOU M, LONG W, LI X. Patterns of synonymous codon usagebias in chloroplast genomes of seed plants [J]. Forest Stud Chin, 2008, 11(4):235-242.

ZHOU M, WEI L, XIA L, 2008. Patterns of synonymous codon usage bias in chloroplast genomes of seed plants[J]. For Ecosy, 10(4): 235-242.

chinaXiv:201809.00100v1